ISSN No. 2278-3091

Volume 2, No.2, March - April 2013
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://warse.org/pdfs/2013/ijatcse01222013.pdf

A Survey on Resource Management in Data Grid

Somayeh Abdi[1], Sayyed Mohsen Hashemi[2]
[*1] Department of Engineering, Eslamabad-E-Gharb Branch, Islamic Azad University, Eslamabde Gharb, Iran
somayeh.abdi@gmail.com

[2]Department of Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran
hashemi@srbiau.ac.ir

ABSTRACT

Nowadays, scientific, business and academic centers submit their large-scale applications (in term of computations or data) into grid environment to be performed. Resource management is important issue in such environment. Resource management divides in resource allocation and data management. Resource allocation refers to resource discovery and resource scheduling, while data management includes replication mechanisms. In data management issue, Data Grid prepares mechanisms for storage management and efficient access to data. In resource allocation issue, for effective performing the applications, greater efficiency and higher utilization of grid resources, tasks scheduling should be done as effectively and data replication mechanism should be applied to reduce data access latency. In this paper, classification of job scheduling and data replication on data grids and their characteristics are being expressed.

Keywords: *data grid, job scheduling, resource discovery, grid information system, data replication.*

1. GRID SYSTEM

Grid is a mechanism that provides the ability of sharing resources and access to the distributed environment as well as virtualization technologies, scheduling techniques, security, and resource management. A real and important issue in Grid systems is sharing coordinated resource and solving dynamic problems in virtual organizations. The concept of resource sharing is not exchanging the files, but direct access to computers, software, data and other resources. This account is strictly controlled and it has accurately and been declared for suppliers and users of resources to what is shared, Who will have access to shared resources, and under what conditions resource sharing will be done. A collection of entities that are defined under common rules, called virtual organizations [1, 2, 3, 4]. There are several types of grid architecture, which are used for several types of commercial and scientific issues. Architecture type will be chosen based on commercial or scientific problem solving, and on the other hand commercial or scientific purposes should also be considered in selecting the grid kind. Some grid architectures have been designed to benefit from the advantages of more computing resources, such as grid computing. Grid computing is hardware and software infrastructure that provides accessible, independent, consistent and inexpensive computational resources. Some of the other grid architectures are designed to provide collaboration between virtual organizations, Such as grid service. Data Grid is the important kind of grid. The data grid is a structure that manages data, storage and network resources that are located in distinct administrative domains with local policies and national policies. Data Grid provides high speed and reliable access to data by determining how to effectively use the data and scheduling of resources [5, 13] .

2. JOB SCHEDULING

Nature of facing with the data / computation in large scale and distributed in grid create challenges. One of the important challenges is how to increase the performance of grid system by effective scheduling tasks. Scheduling policies play a great role in grid for distributing jobs to the appropriate resources. Effective scheduling methods could reduce job execution time and increase the efficiency of the grid. Job scheduling defines as assigning a job to specific resource to be perform. For each job that submitted to grid environment, grid scheduler assigns it to specific resource according to global scheduling policy, and after that, each resource schedules and performs the tasks according to its local policy. The scheduling architectures have different characteristic such as scalability, autonomy, fault tolerance, quality of scheduling decision. Scheduling architectures include three main categories: centralized, hierarchical and distributed scheduling [18, 13]. These architectures have been compared for different characteristics in table 1. In different types of scheduling architecture, the global and local levels in the implementation scheduling policy will vary. To schedule a job to a resource the Grid scheduler obtains information about available resources from Grid information service (GIS). It is called resource discovery. Then, based on the strategy and scheduling algorithm determines the appropriate site for performing the job. In a centralized architecture, the Grid scheduler determines a suitable site for job and delivers it to resource scheduler. The resource scheduler (each resource has several processing elements) assigns the job to a processing element for performing. In this type of scheduling, Scheduling criteria are defined on the resource level and increasing the number of resources can lead to bottlenecks in the grid scheduler. Figure 1 shows sequence Diagram of centralized job scheduling.

Table 1: *Classification of Scheduling Architecture*

Architecture	Scheduler	flexibility	scalability	Interoperability	Autonomy	Scheduling information	Fault tolerance	Scheduling	Project
centralized	grid scheduler	Low	Low	Low	Low	global information	low	optimal	Condor , MSHN
distributed	regional schedulers	High	High	High	High	regional information	High	Sub-optimal	AppLeS, Ninf, Javelin, NetSolve,
hierarchical	global schedulers and regional schedulers	High	High	High	High	global and regional information	High	optimal	Darwin , Nimrod/G

In hierarchical architecture, jobs are submitted to Grid Scheduler and it selects appropriate region for a job, then the job will be delivered to the regional scheduler. The Regional Resource Scheduler assigns the job to the appropriate site in the domain. In this type of scheduling, Scheduling criteria are defined at regional level. In this structure, increasing number of resources and users do not lead to create bottleneck in the Grid scheduler because scheduling decision overhead is distributed between Grid scheduler and regional schedulers. Figure 2 shows sequence Diagram of hierarchical job scheduling. In distributed architecture, there are several domains in data grid and in each domain there is a grid scheduler that is responsible for scheduling the jobs that started in it.

The main objective of scheduling strategies can determine the scheduling algorithm. Scheduling strategies are put in three categories which include: trust-base, market-based and performance-base. In the scheduling strategy based on trust, scheduler selects resources with confidence level of security thus increase the accuracy and reliability of the system. In market-based, scheduler uses economic model to allocate resources. In performance-based, scheduler considers performance of tasks that are scheduled specifically to improve system performance [13]. These performance criteria are minimizing job execution time, increasing the efficiency of computing resources and network resources.

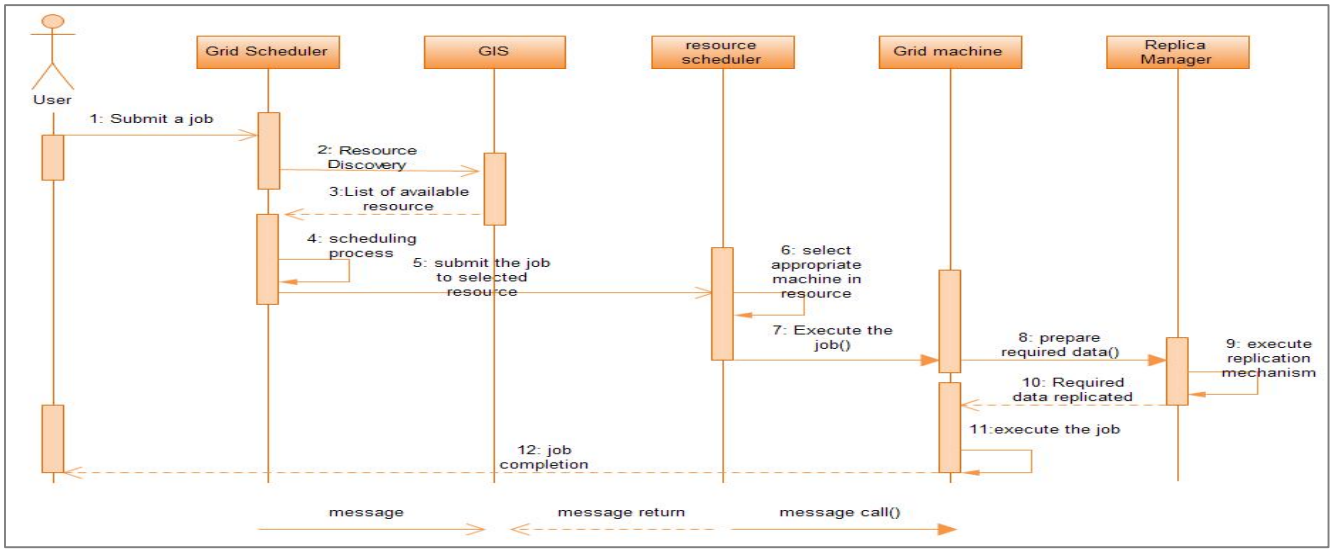

Figure 1: Sequence Diagram of centralized job scheduling

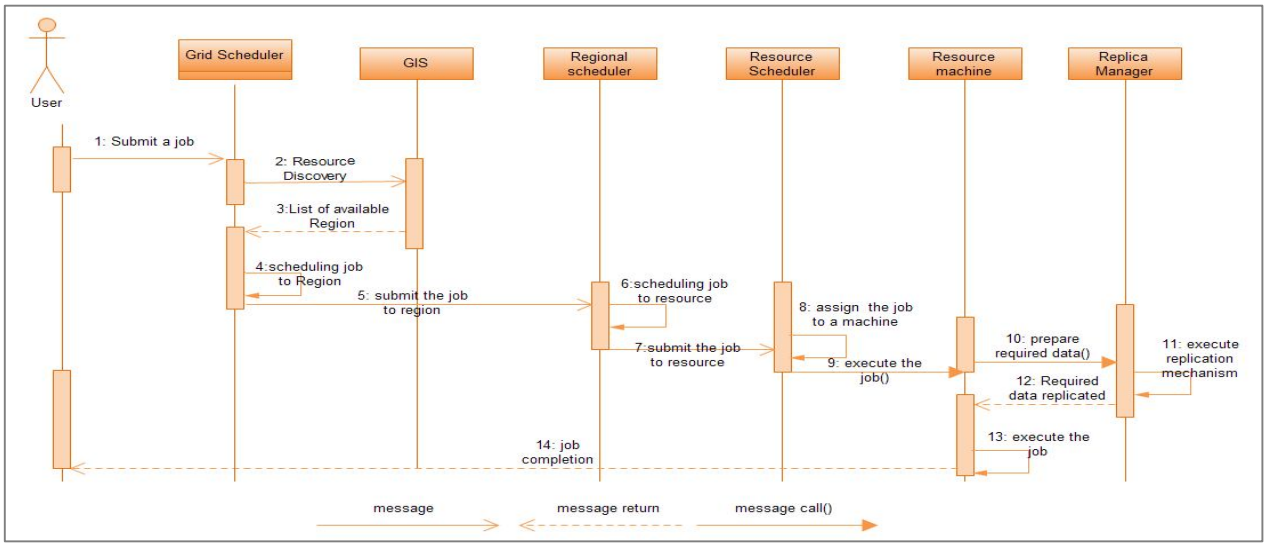

Figure 2: sequence Diagram of hierarchical job scheduling

3. DATA REPLICATION

There are two main reasons for the data replication: reliability and performance. When data replication is multiplied and spread to different parts of a system, the system will reach a higher level of reliability. If the file is corrupted in such a system (or cannot access it), the system will simply switch to other versions that it does not come to interfere in the work. The resolution of such a system is more robust against data corruption.

Another reason is increasing the efficiency of system for data replication. [17] Data Grid Location influence on decision-making and effectiveness data replication is required to do the job scheduling for data intensive applications. Because the data grid will show a distributed solution, its performance under the impact of network latency and bandwidth are available. In data intensive applications, data transfer time is running late due to job requirements. To reduce time of data transferring, data replication is used to improve the performance of the job execution. Replication mechanism determines how files should be repeated; create new versions, when they are available by everywhere [14].

In terms of the dynamics of the Grid environment for data replication mechanism, two types of data replication are: dynamic data replication and static data replication [8]. The static data replication does not consider data access pattern, so changing of data access patterns will reduce the performance of data replication. In contrast, the dynamic data replication approach, considers data access pattern in the Grid environment in replication mechanism. In data replication structure, replication manager manages data transferring between storage resources.

Obviously, different data replication strategies require different access patterns and employing appropriate strategies can save network bandwidth and reduce data access latency.

Thus, replication strategies are combined with data access patterns that include:1) Random Access: Random access to data is not according to locality 2) Data access based on time-locality (Have recently used files that are likely to be requested again). 3) Data access based on time and location locality (Have recently used files that are likely to be requested again in the near site).

Another classification algorithm is based on the structure of the data replication algorithms that are classified into concentrated and distributed categories [13]. In centralized data replication there is a replica manager and several replica servers. Each replica server stores records for data access requests and in specific intervals, all replica servers send information about data request to replica manager to make decision on creating new replicas or removing replicas on replica servers. The decision to create new replica based on global information of data access in grid system is one of the benefits of this approach. But it also has disadvantages; increasing in the number of storage resource or data files in the system caused in decreasing the performance of data management or even it will be impossible. Other disadvantages of this method are to create a bottleneck in replica manager and low fault tolerance in grid system [1]. Figure 3 shows combining centralized data replication and centralized scheduling. After scheduling a task to specific resource, it will request required data from replica server. In this figure, for simplicity, only one replica server is shown. In distributed data replication strategies in grid systems, there are several replica managers in different domains, these components interact with each others to make replica decisions. Distributed data replication advantages are high scalability, without creating bottlenecks in the system and high fault tolerance. Another scalable model has been shown in figure 4. In this structure storage resources act as a replica servers and store required data that are transferred to resources.

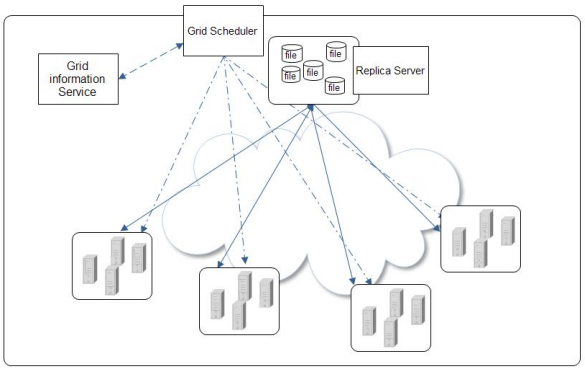

Figure 3: Structure of centralized job scheduling and data replication

This method is useful if the data access patterns are kind of time and location locality based. Figure 4 shows the combination of job scheduling and general model for data replication strategies. In this architecture replica catalogue lists resources that stored each file and interacts with replica manager for efficient data management and updating file indexing information. Different architectures can be combined with scheduling and data replication.

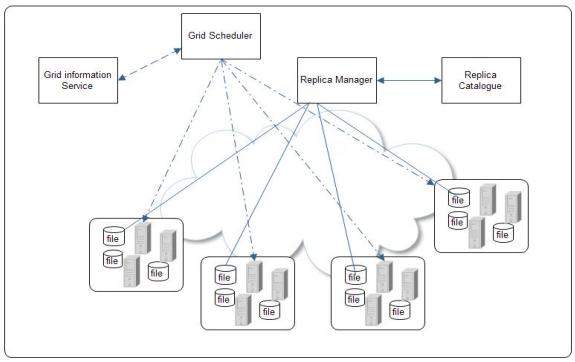

Figure 4: global model of centralized job scheduling and replication

4. CONCLUSIONS

Grid users use the Grid system for solving complex problems in collaborative environment or running applications that require high computing power or processing data in high-volume. The main idea of using the Grid is avoiding establishing specific clusters for such applications and savings the budget. There are several types of grid systems and Grid architecture should be selected based on commercial or scientific problem solving. Grid provides Different type of service to users. In grid computing, processing elements are provided for jobs to run in parallel or distributed, while in data grid, storage resources, data management and efficient access to data are core services offered to users. Nowadays many data grid applications are intensive to data, and data access patterns are different. In Data Grid, data transfer time is the important cause of job execution delay and making congestion in network resources. Therefore by employing appropriate data replication mechanism, scheduling policies and considering the data access pattern, the performance of these applications in Data Grid is increased.

REFERENCES

[1] Avitha Ranganathan and Ian foster,**computation schedulingand data replication algorithms for data grids.**

[2] A. Horri, R. Sepahvand, Gh. Dastghaibyfard, "**A Hierarchical Scheduling and Replication Strategy**", JCSNS International Journal of Computer Science and Network 30 Security, VOL.8 No.8, 2008.

[3] R. Sepahvand ,A. Horri, Gh. Dastghaibyfard, "**A New 3-layer Replication and Scheduling Strategy In Data Grid**". Internatioal Symposium on Telecommunications.

[4] Foster, I., Kesselman, C. and Tuecke, S. **The Anatomy of the Grid: Enabling Scalable Virtual Organizations**. International Journal of High Performance Computing Applications, 15 (3). 200-222.

[5] K. Ranganathan and I. Foster, **Identifying Dynamic Replication Strategies for a High Performance Data Grid.** In Proceedings of the International Grid Computing Workshop, Denver, Colorado, USA, 2001.

[6] I. Foster, K. Ranganathan, **Design and evaluation of dynamic replication strategies for high performance data grids,** in: Proceedings of International Conference on computing in High Energy and Nuclear Physics, Beijing, China, September 2001.

[7] Nhan Nguyen Dang & Sang Boem Lim2, " **Improvement of Data Grid's Performance by Combining Job Scheduling with Dynamic Replication Strategy"**

[8] Nhan Nguyen Dang, Sang Boem Lim2: **Combination of Replication and Scheduling in Data Grids**, IJCSNS International Journal of Computer Science and Network Security, VOL.7 No.3, March 2007.

[9] Ming Tang, Bu-Sung Lee, Xueyan Tang, Chai-Kiat Yeo **." The Impact of Data Replication on Job Scheduling Performance in the Data Grid",** Future Generation Computer Systems, Volume 22, Issue 3, February 2006, Pages 254-268

[10] H. Lamehamedi, Z. Shentu, B. Szymanski**, Simulation of dynamic data replication strategies in Data Grids, in:** Proceedings of 12th Heterogeneous Computing Workshop (HCW'03),April 2001.

[11] M. Carman, F. Zini, L. Serafini, K. Stockinger, **Towards an economybased optimisation of file access and replication on a data grid**, in: Proceedings of 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid, CCGrid 2002, IEEE-CS Press, Berlin, Germany, 2002, pp. 340–345.

[12] Sang-Min Park, Jai-Hoon Kim, Young-Bae Ko: **Dynamic Grid Replication Strategy based on Internet Hierarchy**, Book Series Lecture Notes in Computer Science, Grid and Cooperative omputing book,Publisher Springer, Volume 3033/2004, Pages 838-846.

[13] Klaus Krauter1, Rajkumar Buyya2 and Muthucumaru Maheswaran, **A taxonomy and survey of grid resource management systems for distributed computing,** SOFTWARE—PRACTICE AND EXPERIENCE Softw. Pract. Exper. 2002; 32:135–164 (DOI: 10.1002/spe.432).

[14] Ruay-Shiung Chang, Jih-Sheng Chang, Shin-Yi Lin, **Job scheduling and data replication on data grids,** Future Generation Computer Systems, Volume 23, Issue 7, August 2007, Pages 846-860

[15] Jianhua Jiang, Huifang Ji, **"scheduling algorithm with potential behaviors"**, Journal of Computers, VOL. 3 , NO. 12 , December 2008.

[16] Srikumar Venugopal and Rajkumar Buyya, "**An Economy-based Algorithm for Scheduling Data-Intensive Applications on Global Grids".**

[17] Vidal Martin, Esther pacitti,patrick valduriez. "**Survey of data replication in p2p systems".**

[18] WP01 R E P O R T O N C U R R E N T ,T E C H N O L O G Y Document, "DataGrid".